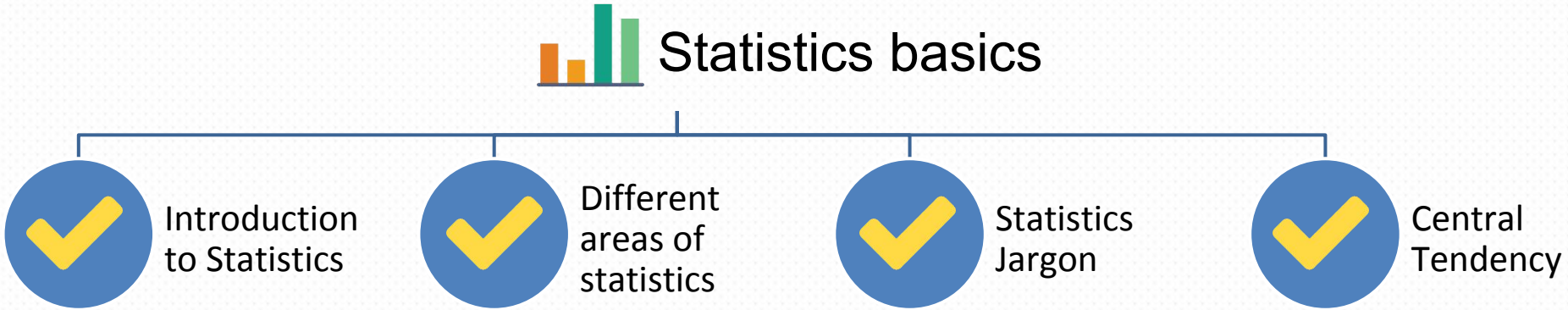
edWisor

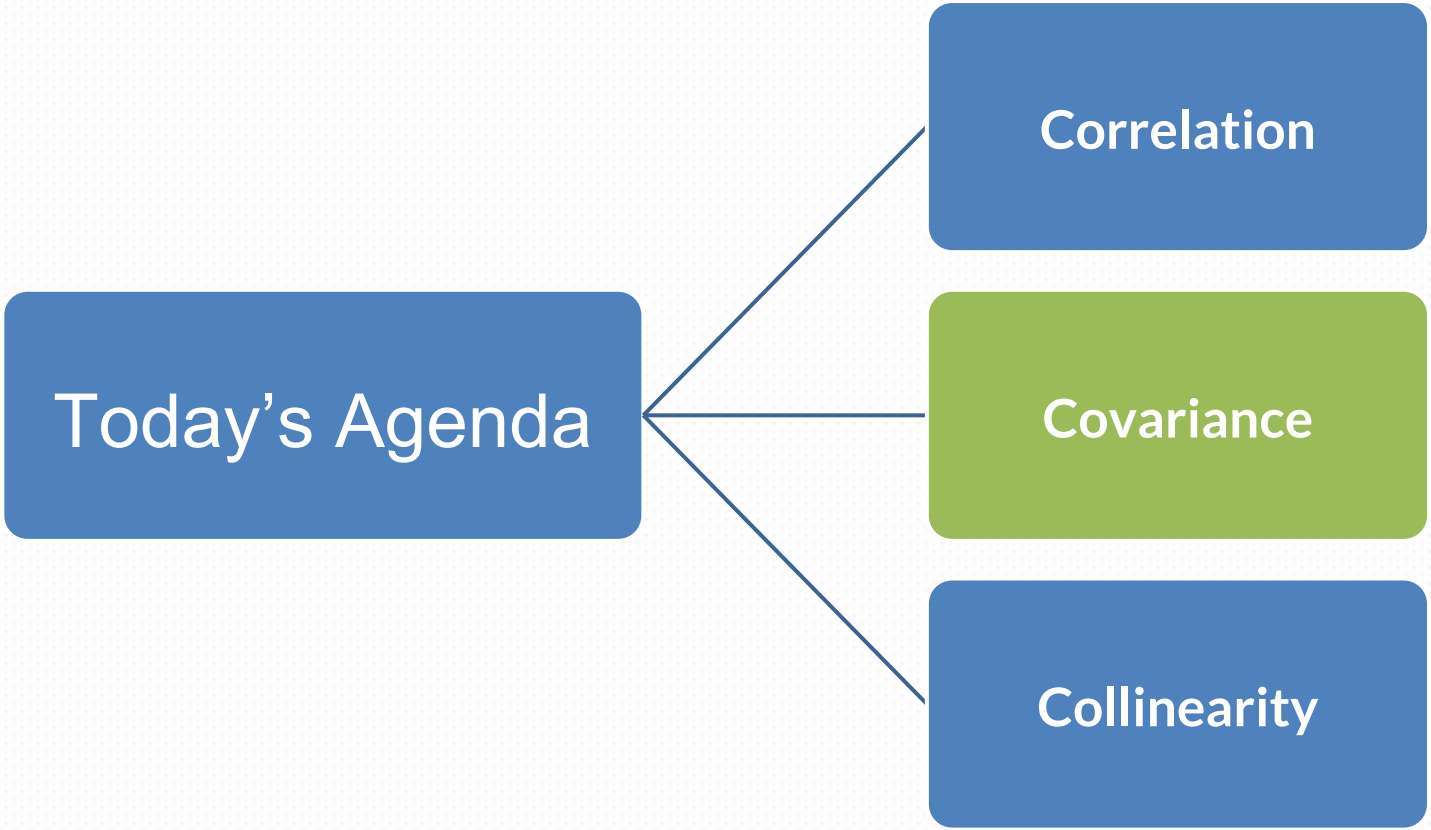
Basic Statistics

- MUQUAYYAR AHMED
DATA SCIENTIST

Agenda We Learnt...



edWisor



Co-variance

- The covariance of two variables x and y in a data sample measures how the two variables are linearly related and it is a measure of how much two random variables change together
- The covariance of two variables x and y can be calculated as

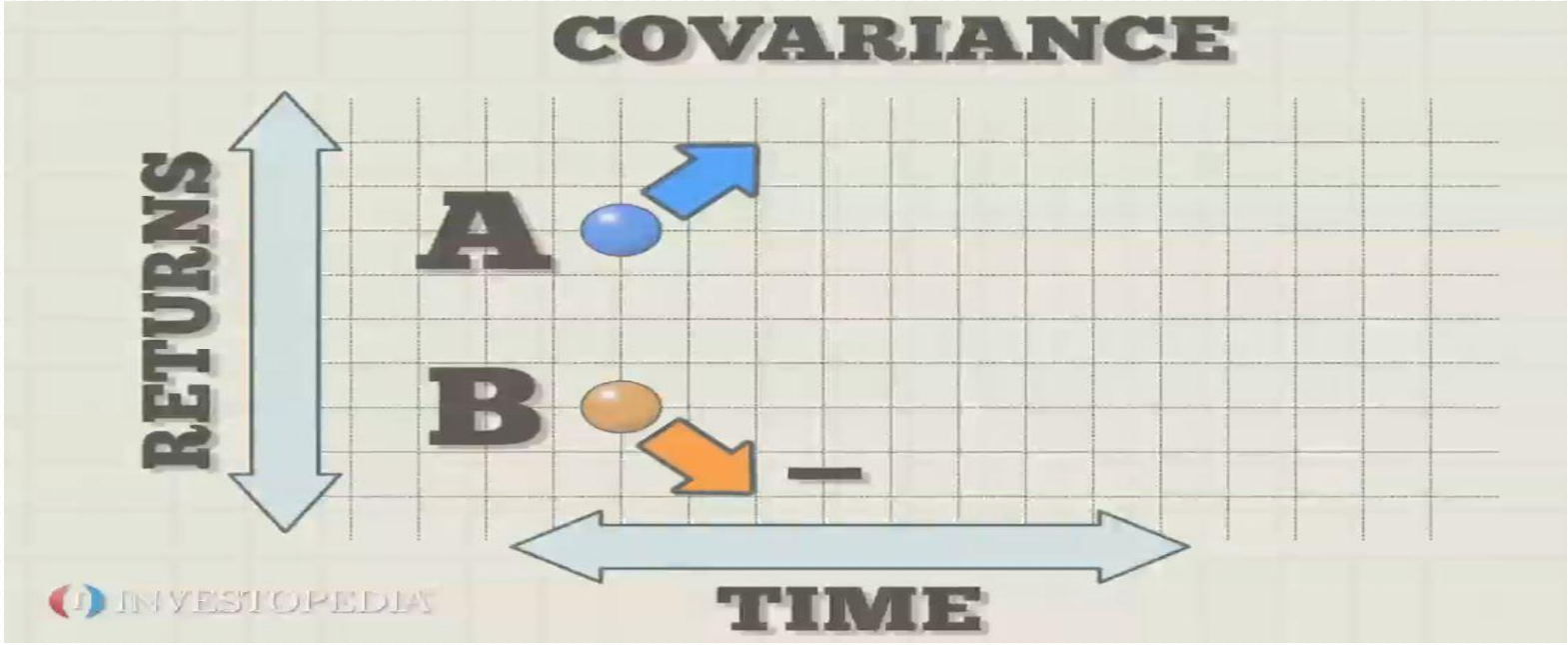
$$COV(X,Y) = \frac{\sum_{i=1}^{n}\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

- A positive covariance would indicates a positive linear relationship between the variables, and a negative covariance would indicate the opposite.

Contd..

Contd..

- A positive covariance would indicates a positive linear relationship between the variables, and a negative covariance would indicate the opposite.



Correlation

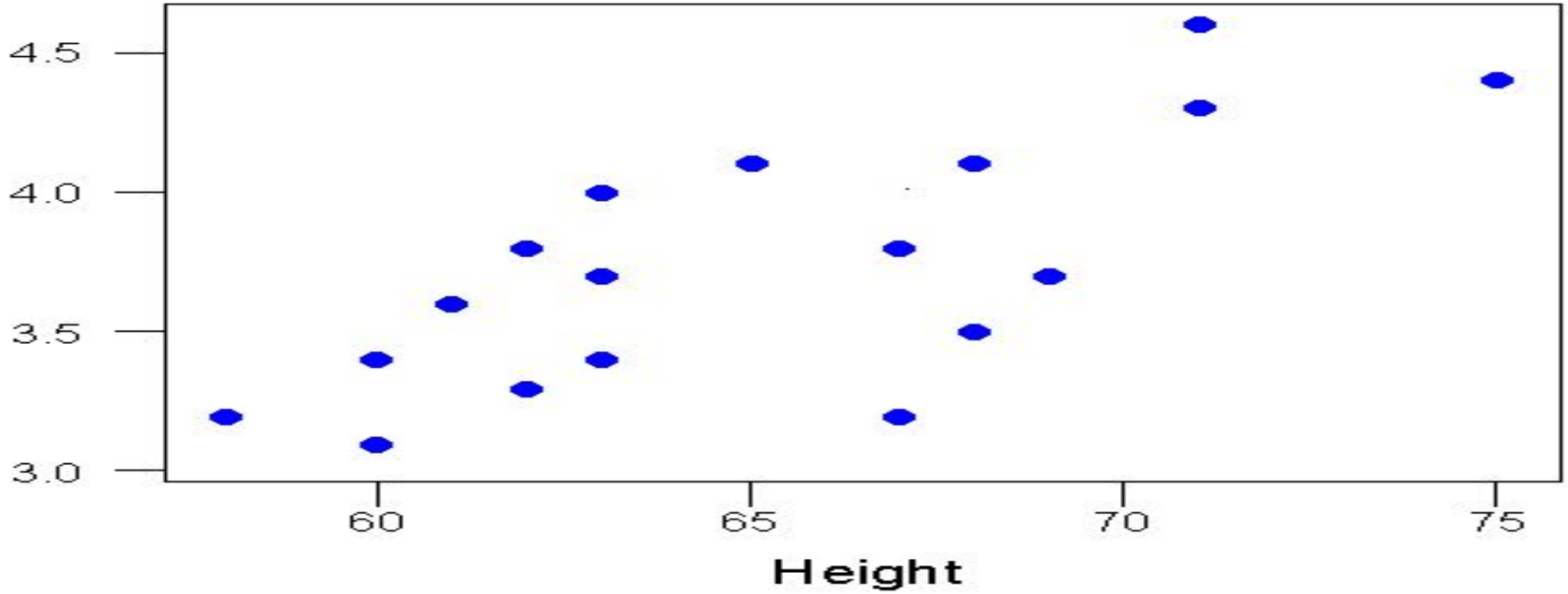
- Correlation is a statistical technique that can show whether and how strongly pairs of variables are related
- It is a scaled version of covariance and values ranges from -1 to +1
- It can be calculated as

$$\mathbf{cor}(X, Y) = \frac{\mathbf{cov}(X, Y)}{\mathbf{sd}(X)\mathbf{sd}(Y)}$$

- Why do we need to look at correlation??

Contd..

- Example of positive correlation



edWisor

Collinearity

- Collinearity or Multicollinearity is the occurrence of several independent variables in a regression model are closely correlated to one another.

- Why is this a problem?
  - Collinearity tends to inflate the variance of at least one estimated regression coefficient.
  - This can cause at least some regression coefficients to have the wrong sign.

- Ways of dealing with collinearity
  - Ignore it. If prediction of y values is the object of your study, then collinearity is not a problem.
  - Get rid of the redundant variables using a variable sélection technique.